looki

清華大學 人工智能学院
College of AI, Tsinghua University

LifeTrail: Benchmarking Multimodal Memory, Personalization, and Proactive Assistance over Long-Horizon Real-World Life Traces

Zixuan Wang[1,2,*] Dayu Yue[2] Yilin Gui[2] Yufeng Zhou[1] Bocong Liu[2,‡] Miao Liu[1,†]

[1]Tsinghua University [2]Looki Tech Limited

wang-zx23@mails.tsinghua.edu.cn miaoliu@mail.tsinghua.edu.cn bocongl@looki.ai

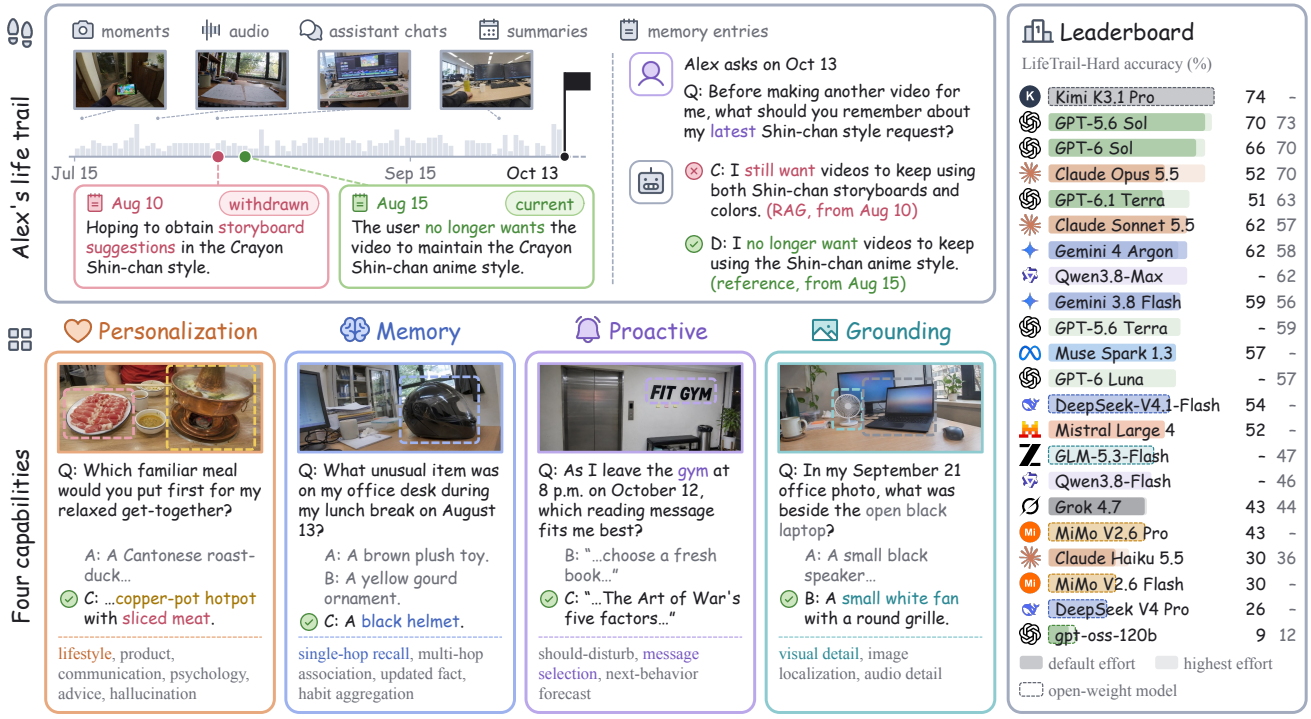


Figure 1. **Overview of LifeTrail.** LifeTrail asks questions over a user's multimodal life records in LifeAtlas to evaluate how well personal assistants understand the user's long-term life. Top left: In a three-month life trail, a request made on Aug 10 is withdrawn on Aug 15, yet when asked about the user's latest request on Oct 13, RAG-based answer models still select the withdrawn one. Bottom left: The four capabilities that LifeTrail evaluates (personalization, memory, proactive assistance, and multimodal grounding), with example questions. Right: Accuracy of 22 models on LifeTrail-Hard with Agentic RAG, at the default reasoning effort of each API (deep bars) and at the highest effort (pale bars). Dashed outlines mark open-weight models.

Abstract

Personal AI assistants must connect a user's everyday experience with their preferences, evolving circumstances, and immediate needs. Existing benchmarks evaluate personalized assistants on chat histories, synthetic profiles, or short recordings rather than on months of real-world experience. We collect LifeAtlas, a dataset of more than 180k multimodal life records, including photos, audio transcripts, locations, and assistant chats, from more than 150 users over up to three months. On top of it, we build LifeTrail, a benchmark of over 1,300 multiple-choice questions spanning 16 tasks in four capabilities: personalization, memory, proactive assistance, and multimodal grounding. An agentic pipeline expands human-written seed questions into grounded questions, and every released question passes cross-model validation and human review. We evaluate memory systems and retrieval methods with three answer models. Memory systems that consolidate the history perform best, yet every setting remains at least 15 points below gold evidence, which makes retrieving the relevant records the main bottleneck of personal agent systems. Agentic RAG improves factual recall over reranked multimodal RAG but lowers accuracy on personalization and updated facts. Proactive inference remains difficult even with gold evidence, and higher reasoning effort brings the largest gains on these questions. We further benchmark 22 frontier models on the challenging LifeTrail-Hard under varying rea-

*Work done during an internship at Looki AI.

†‡ Corresponding author.

soning effort and tool budgets. The strongest model reaches 74% accuracy, and higher reasoning effort helps more than a larger tool budget. The dataset and benchmark will be released upon publication.

1. Introduction

Personal AI assistants are moving from answering isolated requests to acting for a user over weeks and months. Agents built on frontier models now search large personal archives, call tools, and plan multi-step actions [45, 93], and a growing number of memory systems give them persistent state across sessions [14, 61, 63, 87]. Useful assistance in this setting requires recalling events from the user's past, tracking how their preferences change, and deciding when an intervention would help [10, 47, 51]. Such systems appear far faster than benchmarks that test them on real users, and comparing them requires longitudinal records of everyday experience that capture how a user's circumstances and preferences change over time.

Wearable lifelogging cameras capture everyday experience through first-person images and audio. These data can be organized into records of conversations, locations, and recurring activities [22, 23, 27, 91], which reveal how a user's preferences and intentions evolve, including cases where newer information contradicts an earlier request. In Fig. 1 (top left), a user requests cartoon-style storyboards on August 10 and withdraws the request five days later. Two months later, the user asks what to keep in mind before making the next video. With RAG-based retrieval, all three answer models select the withdrawn request. A record that was correct when it was made thus leads to a wrong personalized answer once a later change is overlooked.

Existing benchmarks evaluate different aspects of personal memory and reasoning. Dialogue-memory benchmarks emphasize recall and temporal reasoning over text interaction histories [17, 54, 82], personalization benchmarks often rely on synthetic user profiles [34, 35, 59], and egocentric benchmarks focus on visual understanding of first-person recordings [9, 12, 55, 91]. Most of these benchmarks rely on chat histories, synthetic profiles, or short recordings rather than months of real everyday experience, and they evaluate each capability in isolation without testing how memory, personalization, and proactive assistance interact on the same history.

To close this gap, we collect **LifeAtlas**, a dataset of multimodal life records from more than 150 users who wore a lifelogging camera for up to three months. On top of LifeAtlas, we build **LifeTrail**, a benchmark of over 1,300 multiple-choice questions spanning 16 tasks in four capabilities: *personalization*, *memory*, *proactive assistance*, and *multimodal grounding* (Fig. 3 and Table 1). Every question fixes an anchor time and allows only earlier records. Writing such questions by hand does not scale to months of records from many users, so we automate construction with an agentic pipeline (Fig. 2). Starting from human-written seed questions, it samples task-specific evidence, writes first-person questions grounded in the records, and verifies each candidate with independent agentic checks. Each question that passes then goes through human review: users review the personalization and proactive questions built from their own records, and engineers review the memory and grounding questions.

Our contributions are:

- **Dataset.** LifeAtlas, over 180k real-world multimodal life records from more than 150 users over up to three months (Sec. 3.1), which supports research on memory systems, egocentric understanding, and personal agents.
- **Benchmark.** LifeTrail, 16 tasks across four capabilities that share the same histories and temporal access rule, and the challenging LifeTrail-Hard (Secs. 3.2 and 4). Its agentic data engine expands human-written seed questions and applies to other collections of personal records.
- **Evaluation.** A systematic study of memory systems, retrieval methods, and 22 frontier models under controlled reasoning effort and tool budgets (Sec. 5). It shows that locating the relevant records is the main bottleneck, that agentic retrieval gains factual recall at the cost of personalization, and that proactive inference requires reasoning beyond the retrieved evidence.

2. Related Work

Long-Term Memory. Long-term memory benchmarks test whether an assistant can retain and update information over extended interaction histories, mostly in text dialogue [86, 101]. LoCoMo [54] evaluates recall and temporal reasoning over multi-session conversations, and PerLTQA [17] evaluates them over mixed episodic and semantic memories. LongMemEval [82] adds knowledge updates and abstention, MemoryAgentBench [31] evaluates selective forgetting under incremental input, and LifeDialBench [105] restricts each query to the preceding history. Multimodal benchmarks and systems ground memory in photos, videos, and audio [7, 26, 36, 43, 50]. ATM-Bench [57], the closest to our setting in data and scale, poses 1,038 human-written questions that resolve personal references over about four years of a person's photos, videos, and emails. LifeAtlas instead covers up to three months from more than 150 users, and LifeTrail evaluates preference inference and proactive assistance together with recall on the same histories.

Memory systems differ in how they build memory and retrieve relevant information for a query. MemGPT [61] and MemoryOS [38] manage memory hierarchies, Zep [66] maintains a temporal knowledge graph, Mem0 [14] keeps compact, incrementally updated facts, A-MEM [87] links

Table 1. **Comparison of LifeTrail with related benchmarks.** The columns follow the memory abilities of Guo *et al*. [25]: cross-modal evidence (CE), long-range evidence (LE), state changes (SC), conflicting memories (CF), unanswerable queries (UQ), and memory-grounded action (MA), plus personalization (👤) and proactive assistance (⏰), added by us. ✓, ○, and ✗ mark direct, partial, and no evaluation. Ground truth: answer (A), evidence records (E), timestamps (T), actions or states (S), rubric (R).

Benchmark	Source	Span	Ground Truth	CE	LE	SC	CF	UQ	MA	👤	⏰
Image and conversational memory											
LoCoMo [54]	Hybrid	months	A+E	○	✓	○	✗	✓	✗	✗	✗
Memory-QA [36]	Real	months	A+E	○	✓	✗	✗	✗	✗	✗	✗
ATM-Bench [57]	Real	years	A+E	✓	✓	✓	✓	✓	✗	✓	✗
Mem-Gallery [7]	Synth.	Unknown	A+E	✓	✓	✓	✓	✓	✗	○	✗
MemEye [26]	Hybrid	Unknown	A+E	✓	✓	✓	✗	✗	✗	✗	✗
PersonaMem-v3† [35]	Hybrid	30 days	A+R	✗	✓	✓	○	✓	✓	✓	✓
Egocentric video											
EgoLifeQA [91]	Real	1 week	A+T	✓	✓	○	✗	✗	✗	✓	✗
S-EMBER [78]	Real	Unknown	A+E+T	○	✓	○	✗	✗	✗	✗	✗
EgoMonth [12]	Real	20–120 days	A+T	○	✓	✓	✗	✗	✗	✗	✗
Personal agents											
KnowU-Bench [11]	Hybrid	Unknown	S	✓	✓	✓	✗	✗	✓	✓	✓
MEMORA-Bench [98]	Real	Unknown	A+S	○	✓	○	✗	✓	✓	✓	✗
LifeTrail (ours)	Real	3 months	A+E+T	○	✓	✓	○	✓	○	✓	✓

notes, MemFit [63] indexes verbatim records with segment summaries, and ViLoMem [8] accumulates multimodal semantic memory. Retrieval methods range from text and multimodal embeddings [19, 37, 42, 44, 65] with LLM reranking [73] to graph- and tree-based indexes [3, 18, 28, 41, 68, 88] and agentic search with tools [45, 46, 49, 93–96]. These designs trade compactness against fidelity: LongMemEval reports that fact compression improves multi-session reasoning but lowers overall accuracy on dialogue histories. We evaluate representatives of each design on real lifelogs against a gold-evidence setting, and Table 1 places LifeTrail in the memory-ability framework of [25].

Lifelogging and Egocentric Understanding. Lifelogging studies the capture and retrieval of real-world experience recorded by wearable devices [22, 27, 74], in which egocentric data play an increasingly important role. Egocentric datasets have expanded from kitchen and skilled activities [15, 24] to daily life recorded with multi-sensor wearables [52, 53], and benchmarks have lengthened the horizon and broadened the tasks. Ego4D [23] introduced episodic-memory queries, later extended to question answering [6, 33] and to interactive retrieval with user feedback [72]. EgoSchema [55], HourVideo [9], and general long-video benchmarks [20, 107] extend question answering to long-form footage, where agents search captions or frames [77, 80, 99]. Other work studies the history of a single wearer [29, 78, 81, 85]. EgoLife [91] records six participants sharing one week and evaluates recall, habits, and recommendations, while EgoMonth [12] evaluates spatiotemporal memory over 20–120 days. Other benchmarks target real-time assistance, audio-visual understanding, and web tasks [21, 70, 89, 97].

Personalized LLMs and MLLMs. Personalization methods adapt model parameters, augment inputs with user context, or align outputs with preferences [59, 83, 102]. Adaptation methods learn user-specific visual concepts from images and videos [1, 2, 58, 62, 69, 71, 106]. Among augmentation methods, RAP [30] retrieves concepts from an editable database, and PersonaVLM [59] couples multimodal memory with personality-aware generation. Alignment methods such as Personalized Soups [32] merge models trained for separate preference dimensions. On the evaluation side, text benchmarks condition on personas or user profiles [40, 67, 100] or on preferences stated in conversation [104]; PersonaMem [34] tests responses to evolving synthetic profiles, and LifeVal [92] evaluates the completion of personalized consumer tasks. LifeTrail provides no profile or concept set: models must infer preferences and habits from the recorded history.

Proactive Assistance. A proactive assistant decides whether and when to act and what message to deliver [16], and receptivity to such interruptions has long been studied for mobile notifications [56]. Proactive Agent [51] studies task initiation from observed user activity. Streaming video models decide when to respond [10, 64, 79] and are evaluated on online understanding [48, 60] and response timing [4]. Egocentric task guidance, planning, and anticipation [5, 13, 76, 103] focus on a single recorded activity. ProactEgoMem [47] tests the retention of transient information and object states in egocentric streams, and PersonaMem-v3 [35] evaluates cross-platform recommendation, including when to refrain. KnowU-Bench [11] and ProactiveMobile [39] ask mobile agents to anticipate needs from behavioral histories, and MEMORA [98] pairs egocentric recall with planning. LifeTrail evaluates interruption decisions, message selection, and next-behavior forecasting on real users' lifelogs across diverse daily scenarios, on the same histories as its memory and personalization tasks.

3. Dataset and Benchmark

3.1. The LifeAtlas Dataset

LifeAtlas contains the histories of more than 150 users who wore a lifelogging camera between mid July and mid October 2026. Its basic unit is the *moment*, a continuous activity segment with photos, an audio transcript, a location, and timestamps. In total, LifeAtlas holds over 180k records, including over 95k moments that contain over 250k photos. The records were collected in mainland China under the Personal Information Protection Law (PIPL). All users gave explicit, informed consent to the research use and anonymized release of their records and are identified only by pseudonyms. Because the dataset is released to researchers worldwide, we also follow the EU GDPR and U.S. privacy law and apply the strictest of the three. Before release, we blur the faces of users and bystanders, remove personal identifiers, generalize home and workplace loca-

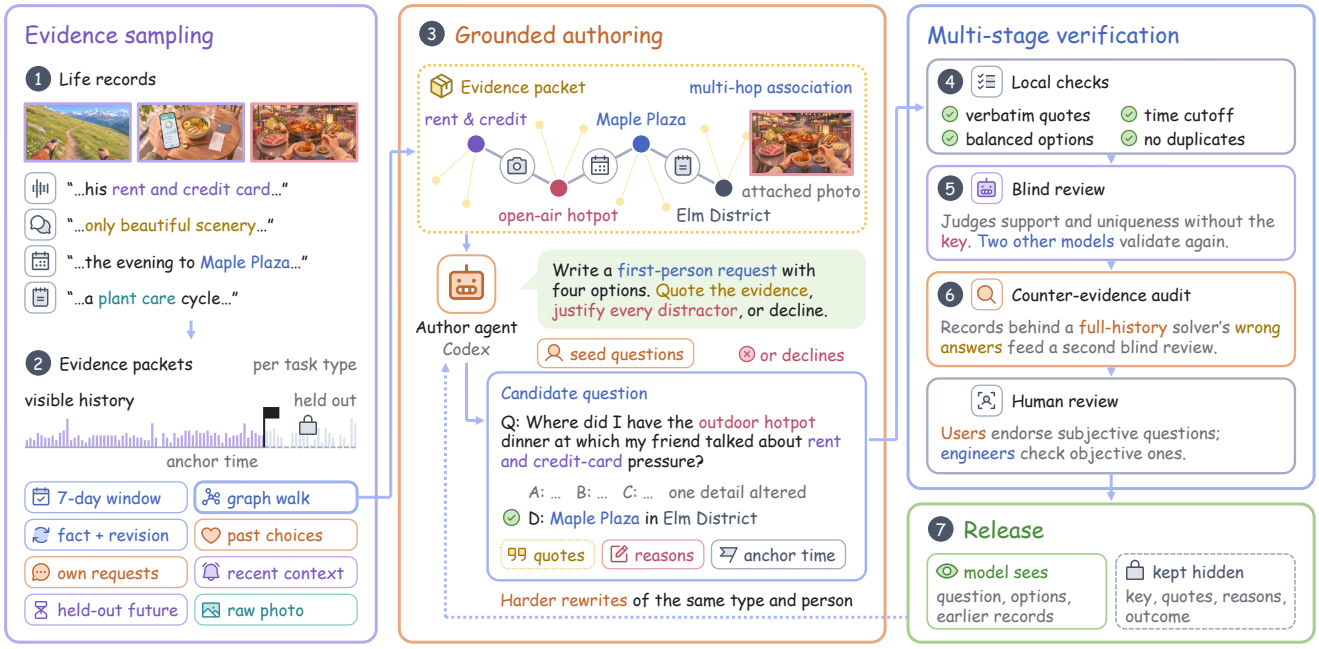


Figure 2. **Overview of the LifeTrail construction pipeline.** Task-specific samplers draw evidence packets from the records before an anchor time (1-2). An author agent then writes a first-person question grounded in the packet, with human-written seed questions as style examples (3). Candidates pass local checks, blind review with cross-model validation, a counter-evidence audit, and human review (4-6). At release, models see only the question, options and earlier records (7).

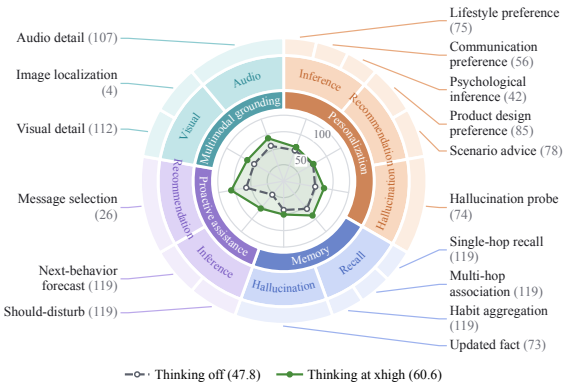


Figure 3. **Task taxonomy of LifeTrail.** The rings show the four capabilities (inner), the nine task groups of Table 2 (middle), and the 16 tasks with their question counts (outer). The radar in the center gives the accuracy of Qwen3.8-Flash in the Agentic RAG evaluation setting on each task group, with thinking off and at the xhigh reasoning effort.

tions, and withhold raw audio.

3.2. Task Definition

Each user u has a history $\mathcal{H}_u = \{r_i\}_{i=1}^{N_u}$, where a record $r_i = (\tau_i, x_i)$ pairs its end time τ_i with multimodal content x_i, such as photos, an audio transcript, a location, or text. A question $q = (u, t_q, x_q, \mathcal{O}_q)$ consists of an anchor time t_q, a first-person request x_q, and an option set $\mathcal{O}_q$ with four options, or two for most should-disturb questions. A model f observes only the visible history $\mathcal{H}_u^{<t_q} = \{r_i \in \mathcal{H}_u : \tau_i < t_q\}$ and selects

$$\hat{o}_q = f\big(x_q, \mathcal{O}_q, \mathcal{H}_u^{<t_q}\big) \in \mathcal{O}_q, \tag{1}$$

which is correct if it equals the reference option o_q^*. The 16 tasks share this interface and differ in how the supporting records E_q determine o_q^* (Fig. 3).

Personalization asks for the option most consistent with the user's preferences and traits θ_u, $o_q^* = \arg\max_{o \in \mathcal{O}_q} s(o; \theta_u)$, where s measures consistency and θ_u is inferred from many records in E_q rather than read from one record. A hallucination probe embeds a preference that no record in $\mathcal{H}_u^{<t_q}$ supports, and o_q^* states that the history does not support it.

Memory asks for a fact stated in the history, $o_q^* = g(E_q)$ with $E_q \subseteq \mathcal{H}_u^{<t_q}$, where E_q is one record for single-hop recall, a chain of records for multi-hop association, and all instances in a window for habit aggregation. For an updated fact, $E_q = \{r_a, r_b\}$ with $\tau_a < \tau_b < t_q$ and r_b revising r_a, so o_q^* follows r_b and a distractor follows r_a.

Proactive Assistance asks for a decision at t_q from the recent context $\mathcal{C}_q = \{r_i : t_q - \delta \le \tau_i < t_q\}$ and the earlier history. Should-disturb decides whether to interrupt at t_q, and message selection chooses the message that fits $\mathcal{C}_q$ and θ_u. Next-behavior forecast asks for the activity $a(r^+)$ of a held-out record r^+ with $\tau^+ > t_q$, which f never observes.

Multimodal Grounding asks for a detail determined by a

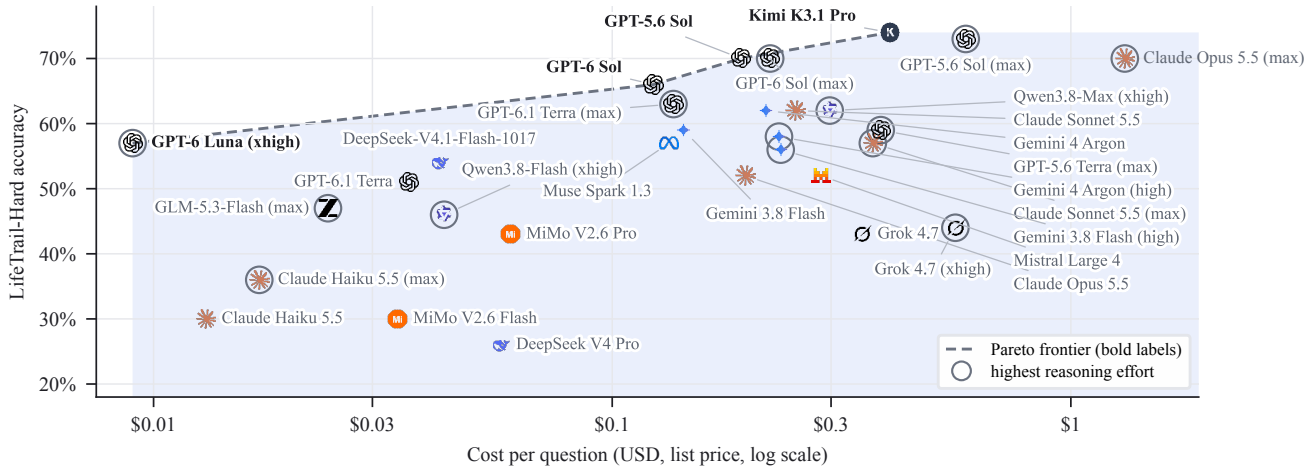


Figure 4. **Accuracy versus average task cost on LifeTrail-Hard.** All models are evaluated with Agentic RAG and a tool-call limit of 15. We omit gpt-oss-120b, which reaches at most 12%. Plain logos mark the default effort of each model and ringed logos its highest effort. The dashed line is the Pareto frontier (bold labels).

photo whose text description is withheld (visual detail), by the record whose photo matches a query photo (image localization), or by an audio transcript (audio detail). Labels for memory, grounding, and forecasting questions follow directly from the records, whereas labels for personalization and the two proactive judgments are inferred from the user's behavior and confirmed by the user (Sec. 4.2).

3.3. Benchmark Characteristics

Table 1 compares LifeTrail with related benchmarks on six memory abilities [25]. Cross-day association tests long-range evidence (LE); revised facts test state changes and conflicting memories (SC, CF); hallucination probes test unanswerable queries (UQ). Proactive tasks extend evaluation toward memory-grounded action (MA), and grounding tasks require visual or transcribed-audio evidence (CE). Because all tasks share the same histories and temporal access rule, capabilities can be compared for the same user. For 533 of the 1,327 questions, the earliest supporting record precedes the question by more than 60 days. Multiple-choice answers make scoring deterministic and comparable across models without a language-model judge.

3.4. Supported Research

Beyond this evaluation, the timestamped records and revised facts support research on *memory systems*: how memories are built, updated, and forgotten, including time-aware retrieval and streaming evaluation. For *personalization*, preferences are inferred from recorded behavior instead of a given profile. For *proactive assistance*, held-out future records label forecasts, and user-endorsed decisions label when to interrupt. For *agentic search*, LifeTrail-Hard measures how agents trade accuracy against reasoning effort, tool budget, and cost.

4. Agentic Data Engine

Unchecked generation produces questions that are unsupported, ambiguous, or answerable without the user's history. To avoid these failures at scale, we design an agentic data engine that grounds every question in sampled evidence and filters every candidate with independent agentic checks before human review (Fig. 2).

4.1. Grounded Construction by Evidence Sampling

Construction starts from seed questions that we write in the first person, each with a note on what separates the correct option from its distractors. Task-specific samplers then draw evidence packets around an anchor time, such as a chain of linked records for multi-hop association or a statement and its revision for updated facts; proactive packets follow ContextAgent [90]. A Codex agent with GPT-6.1 Sol takes the seeds as examples of style and reasoning only and writes a new first-person question from each packet, with the correct option and verbatim evidence quotes. Every distractor is plausible for the same user and comes with a rationale that cites the evidence against it.

4.2. Multi-Stage Verification

Local Checks enforce the question format, verbatim quotes, and the anchor-time rule.

Blind Review and Cross-Model Validation. A fresh GPT-6.1 Sol session receives the question, options, and permitted records, without the reference answer. It must select the author's option and confirm that the answer is supported, unique, and not leaked. Two other models, GPT-6-Astra and GLM-5.3-Flash, then cross-validate every question.

Counter-Evidence Audit. Qwen3.8-Flash answers each surviving question from the user's full visible history.

Table 2. **Main results on LifeTrail.** Accuracy (%) ranges from no evidence (lower bound) to gold evidence (upper bound). Memory methods build a query-agnostic memory, and RAG methods retrieve records for each query (MM: multimodal). Columns are the task groups of Fig. 3, where memory Hallu. holds the updated facts. Qwen and DeepSeek are Qwen3.8-Flash and DeepSeek-V4.1-Flash without thinking, and Luna is GPT-6 Luna with medium thinking. Shaded cells are the best in each column between the two bounds for Qwen (purple), DeepSeek (blue), and Luna (green). ▲/▼ mark Agentic RAG at least 8 points above or below MM RAG w/ rerank.

	Method	Answerer	Personalization			Memory		Proactive		Grounding		Overall ↑
			Infer.	Rec.	Hallu.	Recall	Hallu.	Infer.	Rec.	Visual	Audio	
Lower	No evidence	Qwen	50.3	46.0	71.6	34.7	45.2	17.6	53.8	33.6	43.0	38.7
		DeepSeek	49.7	46.6	52.7	33.9	52.1	20.6	53.8	38.8	36.4	38.2
		Luna	53.2	48.5	58.1	37.8	53.4	41.2	61.5	34.5	41.1	44.2
Memory	A-MEM [87]	Qwen	**66.6**	**62.6**	59.2	**58.4**	**65.8**	19.2	**68.5**	44.7	**78.7**	**54.0**
		DeepSeek	60.9	55.7	55.1	51.1	62.8	29.1	68.0	49.1	**73.7**	51.9
		Luna	**78.2**	**71.8**	**86.9**	**67.1**	69.2	53.6	89.9	**71.3**	**86.0**	**70.2**
	Mem0 [14]	Qwen	64.8	57.6	60.9	55.9	60.4	16.6	62.2	44.1	76.8	51.5
		DeepSeek	67.8	**68.9**	**60.9**	**58.7**	64.0	26.7	**82.4**	46.1	69.0	**56.0**
		Luna	75.0	70.4	82.4	66.6	69.8	**58.6**	**91.0**	51.5	80.9	68.1
	MemFit [63]	Qwen	52.0	57.1	52.7	43.4	53.4	17.6	57.7	44.8	63.6	44.7
		DeepSeek	59.0	56.4	56.8	47.3	61.6	29.4	69.2	55.2	64.5	50.6
		Luna	65.3	62.6	85.1	54.6	65.8	56.7	88.5	60.3	72.9	62.3
RAG	Naive RAG	Qwen	65.9	60.7	59.5	51.0	61.6	16.4	57.7	48.3	64.5	50.0
		DeepSeek	69.9	61.3	60.8	49.9	61.6	30.3	69.2	53.4	57.9	53.0
		Luna	75.1	65.0	74.3	61.1	72.6	52.5	76.9	69.8	73.8	65.3
	MM RAG w/o rerank	Qwen	64.2	55.8	63.5	44.0	58.9	18.1	61.5	46.6	43.9	45.9
		DeepSeek	66.5	58.3	60.8	42.3	63.0	30.3	65.4	45.7	44.9	48.4
		Luna	68.8	62.6	75.7	49.6	60.3	49.6	69.2	47.4	52.3	56.1
	MM RAG w/ rerank	Qwen	65.3	62.0	**64.9**	44.0	64.4	18.5	65.4	39.7	57.9	47.9
		DeepSeek	**71.1**	61.3	59.5	43.4	**67.1**	29.0	73.1	44.8	49.5	50.0
		Luna	73.4	65.0	82.4	47.9	**74.0**	51.7	76.9	50.0	57.0	58.9
	Agentic RAG	Qwen	49.7 ▼	51.5 ▼	48.6 ▼	54.9 ▲	43.8 ▼	**26.9** ▲	57.7	**51.7** ▲	57.0	47.8
		DeepSeek	57.8 ▼	52.1 ▼	58.1	55.5 ▲	46.6 ▼	**31.1**	65.4	**63.8** ▲	64.5 ▲	52.3
		Luna	52.0 ▼	43.6 ▼	60.8 ▼	59.7 ▲	38.4 ▼	34.9 ▼	61.5 ▼	61.2 ▲	65.4 ▲	51.8
Upper	Gold evidence	Qwen	87.9	92.0	82.4	77.6	91.8	27.7	88.5	80.2	94.4	74.6
		DeepSeek	87.3	87.7	85.1	74.2	93.2	33.6	88.5	82.8	92.5	74.5
		Luna	90.2	93.9	91.9	93.3	91.8	54.2	96.2	91.4	93.5	85.7

Records cited in its wrong answers are added as counter-evidence to a second blind review, which rejects the question if another option becomes defensible.

Human Review. Every released question passes human review. Personalization and proactive questions are subjective, and the user is the only authority on the answer. Each user reviews the questions built from their own records in a questionnaire that offers a rejection option, and we keep only those that the user endorses. Memory and grounding questions are objective, so in-house engineers check each answer against the cited records and remove any question that is wrong or ambiguous.

4.3. Question Selection and LifeTrail-Hard

From over 5k questions accepted by the automated stages and their harder rewrites, the final selection balances user and task coverage and favors questions that GPT-6.1 Sol answers incorrectly in a difficulty screening, which yields 1,327 questions from 154 users. Because GPT-6.1 Sol writes and screens the questions, the benchmark is biased against it, and we exclude it from evaluation. For the frontier-model study, **LifeTrail-Hard** contains 100 questions from 57 users that cover all 16 tasks and are selected by their error rate across screening settings. With gold evidence, Qwen3.8-Flash answers all 100 questions correctly, so their difficulty lies in finding and combining the evidence rather than in the labels.

5. Experiments

We systematically compare memory and retrieval settings on the 1,327 questions of LifeTrail and benchmark mainstream frontier models on the 100 questions of LifeTrail-Hard under varying reasoning effort and tool budgets to understand how test-time scaling affects them.

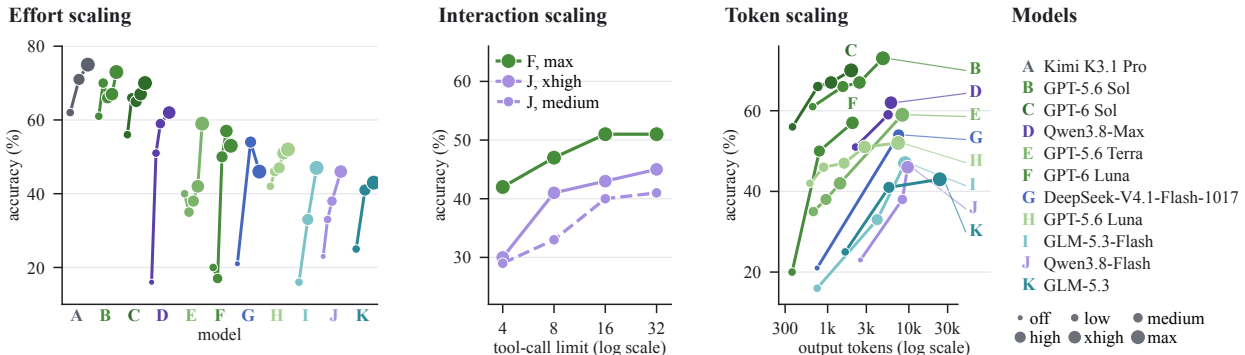


Figure 5. **Test-time scaling on LifeTrail-Hard.** All models use Agentic RAG. Letters identify the models listed on the right, and larger dots mean higher reasoning effort. Left: Each curve shows the accuracy of one model from its lowest to its highest effort. Middle: Accuracy as the tool-call limit grows from 4 to 32. Right: Accuracy against output tokens per question.

5.1. Experimental Setup

Evidence Settings. *No evidence* supplies only the question and options, and *gold evidence* supplies the reference records. Between these bounds, *memory methods* are query-agnostic: A-MEM [87], Mem0 [14], and MemFit [63] build a memory from the visible records before any question is asked. *RAG methods* are query-aware and retrieve the original records for each question. Naive RAG returns ten records ranked by text and photo-caption embeddings, with their photos, Multimodal RAG uses fused text-image embeddings [44] without or with reranking, and Agentic RAG searches moments and opens records within 15 tool calls.

Models and Scoring. The full 1,327-question evaluation uses Qwen3.8-Flash and DeepSeek-V4.1-Flash without thinking, and GPT-6 Luna with medium thinking. On LifeTrail-Hard, we compare 22 models from 11 providers at their default and highest reasoning efforts (Fig. 1, right) and report their cost, following Collaborative Intelligence [84]. Empty or unparseable responses count as wrong. Because most should-disturb questions have two options, the chance level of proactive inference is 36.1%, compared with 25% for the other task groups.

5.2. Main Results

Evidence Access. For each answer model, the highest overall accuracy comes from A-MEM or Mem0, yet every memory and RAG method stays at least 15 points below gold evidence (Table 2). The same answer models answer most questions correctly once the reference records are supplied, so the main difficulty of LifeTrail lies in locating the relevant records within months of history rather than in answering from them. Among the RAG methods, Naive RAG passes about twice as much context as Multimodal RAG with reranking and reaches the highest overall accuracy for all three answer models. Broad coverage of the history matters more than finer ranking of a few records.

Dependence on the Answer Model. GPT-6 Luna with medium thinking exceeds both flash models by 8 to 16 points overall under every memory method and non-agentic RAG method, while the three models differ by at most 6 points without evidence, so a stronger answer model gains by using the supplied evidence better. Agentic RAG is the only method on which Luna does not lead, 18 points below its own A-MEM result. Searching for evidence is a different skill from answering with it and improves with higher reasoning effort (Sec. 5.3).

Query-Agnostic Memory versus Query-Aware RAG. The best memory method exceeds Agentic RAG by 4 to 18 points overall, and the two families lead on different capabilities. Agentic RAG improves memory recall and visual grounding over Multimodal RAG with reranking for all three answer models, but it lowers accuracy on personalization and on updated facts, the memory hallucination group (▲/▼ in Table 2). On updated facts, A-MEM and Mem0 exceed it by 16 to 31 points. The difference follows from when the history is organized. A recall or grounding question is answered by a few records that match its keywords, whereas a preference is expressed across many records and an updated fact requires the later of two conflicting records (Sec. 3.2). A-MEM links related notes and Mem0 merges each new fact into the existing ones while the memory is built, so the aggregation and the update are complete before the question arrives, whereas the agent has to do both within its tool budget. Fact compression lowers overall accuracy on dialogue histories [82], but it helps on lifelogs, which spread one preference or habit over many short records. The original records remain useful: Naive RAG stays within 5 points of the best memory method.

Reasoning Beyond Evidence. Proactive inference, which covers should-disturb and next-behavior forecast, is the only task group on which gold evidence leaves accuracy low: even with gold evidence, Qwen3.8-Flash and DeepSeek-V4.1-Flash stay below its 36.1% chance level, and only GPT-6 Luna with medium thinking exceeds it. These questions ask what the user will do or accept next, which no record states, so the model must infer it from the

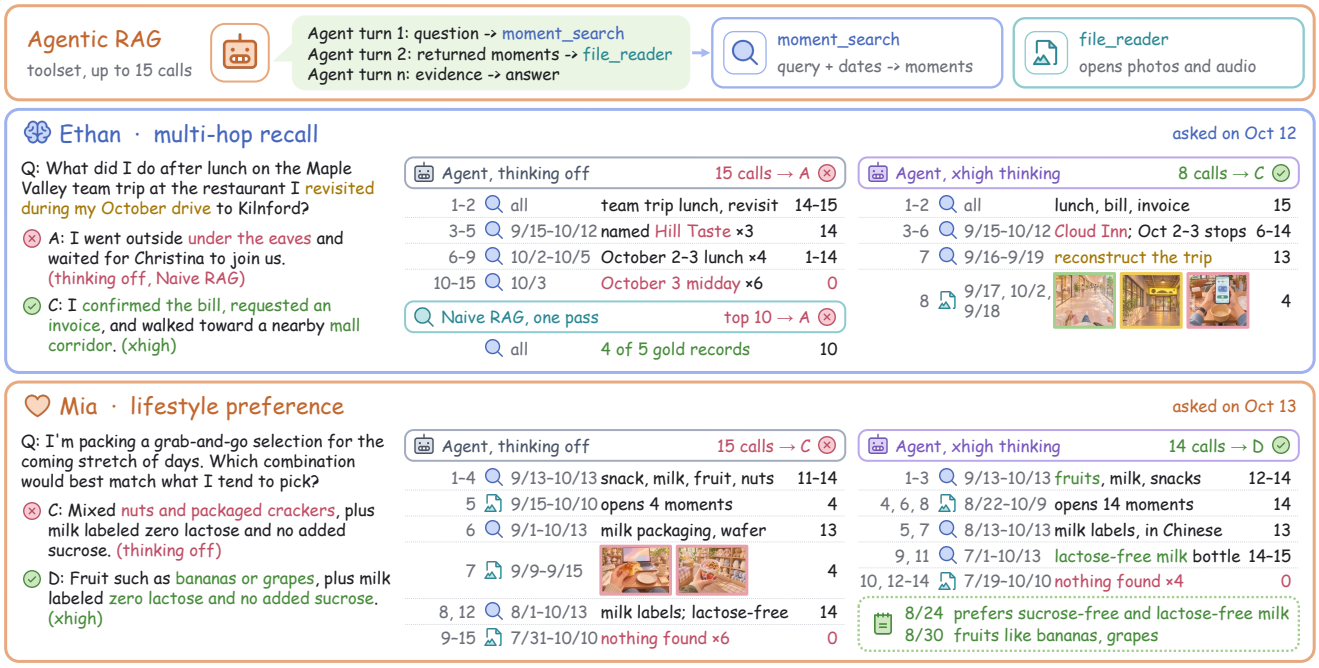


Figure 6. **Visual examples of Agentic RAG on LifeTrail.** Traces of the Qwen3.8-Flash agent with thinking off and at xhigh. Top: Multi-hop recall. Without thinking, the agent spends nine of its 15 calls on an unrecorded period, whereas at xhigh it reconstructs the trip from two connecting moments. Bottom: Lifestyle preference. The preference is stated only in memory entries outside the search index, and only the xhigh agent opens enough moments to recover it.

user's current situation and habits. Reasoning effort supports this inference. On the full benchmark, thinking at xhigh raises the proactive accuracy of Qwen3.8-Flash with Agentic RAG by 26.9 points, the largest gain of any capability, and improves every task group (Fig. 3).

5.3. Additional Analysis

Answer Changes with Evidence. Following Guo *et al*. [25], we compare the answers of the same model with no evidence, in a given setting, and with gold evidence. Compared with Multimodal RAG with reranking, Agentic RAG overturns two to three times as many personalization answers that were correct without evidence, while the share of errors that gold evidence would recover stays nearly unchanged. Its personalization loss thus comes from how the retrieved records are weighed rather than from missing evidence. On proactive questions, errors that gold evidence would recover stay below 14% in every setting, so better retrieval alone does not close this gap.

Effort and Tool Budget. The strongest model on LifeTrail-Hard, the open-weight Kimi K3.1 Pro, answers 74% of the questions (Fig. 1, right). As in MCP-Persona [75], accuracy does not follow cost (Fig. 4): Claude Opus 5.5 at max effort costs about seven times as much as GPT-5.6 Sol at medium effort for the same accuracy. Higher reasoning effort raises accuracy for most models, while raising the tool-call limit from 16 to 32 changes accuracy by at most two points (Fig. 5), so the agent is limited by reasoning rather than by its search budget. With thinking at xhigh, Qwen3.8-Flash halves its tool calls on the full benchmark, and its traces show that thinking makes the search more selective rather than longer (Fig. 6). Higher effort also turns unanswered questions into correct answers: from low to xhigh effort, GPT-6 Luna answers 37 more questions correctly, while its wrong choices stay at 34 and 33.

6. Conclusion

We present LifeAtlas, a dataset of real-world multimodal life records from more than 150 users over up to three months, and LifeTrail, a benchmark of 1,327 multiple-choice questions on personalization, memory, proactive assistance, and multimodal grounding, together with the more challenging LifeTrail-Hard. Experiments with memory methods, RAG methods, and mainstream frontier models under controlled reasoning effort and tool budgets yield three findings. Retrieving the relevant records from months of history is the main bottleneck. Query-agnostic memory outperforms agentic retrieval on personalization and updated facts, while agentic retrieval improves factual recall. Proactive assistance requires reasoning beyond the gold evidence. By releasing LifeAtlas and LifeTrail under strict anonymization, we provide the community with longitudinal records of real lives and an effective evaluation platform for future research on personal intelligence.

LLM Usage Statement

Large language models are part of the method and the object of study in this work. In data construction, GPT-6.1 Sol in a Codex agent writes candidate questions from sampled evidence and performs the blind reviews, Qwen3.8-Flash performs the counter-evidence audit, GPT-6-Astra and GLM-5.3-Flash cross-validate the questions, and language models translate the records into English and audit each translation. Every released question passes human review, and GPT-6.1 Sol is excluded from evaluation. The photos shown in our figures are generated images that reproduce the recorded scenes, so no photo of a real user or bystander appears in the paper. We used coding assistants (Claude Code and Codex) to assist in implementing the pipeline and the figure scripts. We also used language-model assistants to edit the text and to check its consistency with the tables and figures. The authors reviewed all generated code and text, verified every number against the experiment logs, and take full responsibility for the content of the paper.

References

[1] Yuval Alaluf, Elad Richardson, Sergey Tulyakov, Kfir Aberman, and Daniel Cohen-Or. MyVLM: Personalizing VLMs for user-specific queries. In *ECCV*, 2024. 3

[2] Ruichuan An, Sihan Yang, Renrui Zhang, Ming Lu, Tianyi Jiang, Kai Zeng, Yulin Luo, Jiajun Cao, Hao Liang, Ying Chen, Qi She, Shanghang Zhang, and Wentao Zhang. MC-LLaVA: Multi-Concept personalized Vision-Language model. *arXiv preprint arXiv:2411.11706*, 2024. 3

[3] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *ICLR*, 2024. 3

[4] Shehreen Azad, Vibhav Vineet, and Yogesh S Rawat. StreamReady: Learning what to answer and when in long streaming videos. In *CVPR*, pages 40494–40504, 2026. 3

[5] Yuwei Bao, Keunwoo Peter Yu, Yichi Zhang, Shane Storks, Itamar Bar-Yossef, Alexander De La Iglesia, Megan Su, Xiao Lin Zheng, and Joyce Chai. Can foundation models watch, talk and guide you step by step to make a cake? In *Findings of the Association for Computational Linguistics: EMNLP*, 2023. 3

[6] Leonard Bärmann and Alex Waibel. Where did I leave my keys? – episodic-memory-based question answering on egocentric videos. In *CVPRW*, pages 1559–1567, 2022. 3

[7] Yuanchen Bei, Tianxin Wei, Xuying Ning, Yanjun Zhao, Zhining Liu, Xiao Lin, Yada Zhu, Hendrik Hamann, Jingrui He, and Hanghang Tong. Mem-Gallery: Benchmarking multimodal long-term conversational memory for MLLM agents. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 40750–40784, 2026. 2, 3

[8] Weihao Bo, Shan Zhang, Yanpeng Sun, Jingjing Wu, Qunyi Xie, Xiao Tan, Kunbin Chen, Wei He, Xiaofan Li, Na Zhao, Jingdong Wang, and Zechao Li. ViLoMem: Agentic learner with grow-and-refine multimodal semantic memory. In *CVPR*, pages 5476–5486, 2026. 3

[9] Keshigeyan Chandrasegaran, Agrim Gupta, Lea M. Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. HourVideo: 1-hour video-language understanding. In *NeurIPS*, 2024. 2, 3

[10] Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. VideoLLM-online: Online video large language model for streaming video. In *CVPR*, 2024. 2, 3

[11] Tongbo Chen, Zhengxi Lu, Zhan Xu, Guocheng Shao, Shaohan Zhao, et al. KnowU-Bench: Towards interactive, proactive, and personalized mobile agent evaluation. *arXiv preprint arXiv:2604.08455*, 2026. 3

[12] Weitao Chen, Jiaxin Hu, Tianyidan Xie, Li Yang, Yuyi Qian, et al. EgoMonth: A month-level egocentric video benchmark for long-term spatiotemporal memory. *arXiv preprint arXiv:2608.13113*, 2026. 2, 3

[13] Yi Chen, Yuying Ge, Yixiao Ge, Mingyu Ding, Bohao Li, Rui Wang, Ruifeng Xu, Ying Shan, and Xihui Liu. EgoPlan-Bench: Benchmarking multimodal large language models for Human-Level planning. *IJCV*, 134(3):118, 2026. 3

[14] Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. In *European Conference on Artificial Intelligence (ECAI)*. IOS Press, 2025. 2, 6, 7

[15] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100. *IJCV*, 130(1):33–55, 2022. 3

[16] Yang Deng, Wenqiang Lei, Wai Lam, and Tat-Seng Chua. A survey on proactive dialogue systems: Problems, methods, and prospects. In *IJCAI*, 2023. 3

[17] Yiming Du, Hongru Wang, Zhengyi Zhao, Bin Liang, Baojun Wang, Wanjun Zhong, Zezhong Wang, and Kam-Fai Wong. PerLTQA: A personal long-term memory dataset for memory classification, retrieval, and fusion in question answering. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 152–164, 2024. 2

[18] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph RAG approach to Query-Focused summarization. *arXiv preprint arXiv:2404.16130*, 2024. 3

[19] Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. ColPali: Efficient document retrieval with vision language models. In *ICLR*, 2025. 3

[20] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-MME: The First-Ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *CVPR*, 2025. 3

[21] Hengjian Gao, Kaiwei Zhang, Shibo Wang, Mingjie Chen, Qihang Cao, Xianfeng Wang, Yucheng Zhu, Xiongkuo Min, Wei Sun, Dandan Zhu, and Guangtao Zhai. LifeEval: A multimodal benchmark for assistive AI in egocentric daily life tasks. In *CVPR*, pages 32892–32902, 2026. 3

[22] Jim Gemmell, Gordon Bell, and Roger Lueder. MyLifeBits: A personal database for everything. *Communications of the ACM*, 49(1):88–95, 2006. 2, 3

[23] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022. 2, 3

[24] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-Exo4D: Understanding skilled human activity from first- and Third-Person perspectives. In *CVPR*, 2024. 3

[25] Minghao Guo, Qingyue Jiao, Xiujin Liu, Jianhong Pang, Ruimeng Yang, et al. Seeing, maintaining, and learning: The evolution of multimodal agent memory. *arXiv preprint arXiv:2610.19366*, 2026. 3, 5, 8

[26] Minghao Guo, Qingyue Jiao, Zeru Shi, Yihao Quan, Boxuan Zhang, et al. MemEye: A visual-centric evaluation framework for multimodal agent memory. *arXiv preprint arXiv:2605.15128*, 2026. 2, 3

[27] Cathal Gurrin, Alan F. Smeaton, and Aiden R. Doherty. LifeLogging: Personal big data. *Foundations and Trends in Information Retrieval*, 8(1):1–125, 2014. 2, 3

[28] Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. HippoRAG: Neurobiologically inspired Long-Term memory for large language models. In *NeurIPS*, 2024. 3

[29] Tengda Han, Sayna Ebrahimi, Dilara Gokay, Li Yang Ku, Maks Ovsjanikov, Iva Babukova, Daniel Zoran, Viorica Patraucean, Joao Carreira, Andrew Zisserman, and Dima Damen. Unique lives, shared world: Learning from single-life videos. In *CVPR*, pages 24021–24030, 2026. 3

[30] Haoran Hao, Jiaming Han, Changsheng Li, Yu-Feng Li, and Xiangyu Yue. RAP: Retrieval-augmented personalization for multimodal large language models. In *CVPR*, 2025. 3

[31] Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in LLM agents via incremental Multi-Turn interactions. In *ICLR*, 2026. 2

[32] Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*, 2023. 3

[33] Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. EgoTaskQA: Understanding human tasks in egocentric videos. In *NeurIPS*, 2022. 3

[34] Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo J. Taylor, and Dan Roth. Know me, respond to me: Benchmarking LLMs for dynamic user profiling and personalized responses at scale. In *Conference on Language Modeling (COLM)*, 2025. 2, 3

[35] Bowen Jiang, Yuan Yuan, Zhuoqun Hao, Yuchen Liu, Maohao Shen, Sihao Chen, Gregory Wornell, Chris Callison-Burch, Lyle Ungar, Dan Roth, Iordanis Fostiropoulos, Qi Guo, Xiangjun Fan, Camillo J. Taylor, and Hanchao Yu. PersonaMem-v3: Toward omni-platform personal intelligence for holistic user understanding, recommendation, and agentic tasks. *arXiv preprint arXiv:2608.21381*, 2026. 2, 3

[36] Hongda Jiang, Xinyuan Zhang, Siddhant Garg, Rishab Arora, Shiun-Zu Kuo, Jiayang Xu, Aaron Colak, and Xin Luna Dong. Memory-QA: Answering recall questions based on multimodal memories. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 24244–24266, 2025. 2, 3

[37] Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. VLM2Vec: Training Vision-Language models for massive multimodal embedding tasks. In *ICLR*, 2025. 3

[38] Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. Memory OS of AI agent. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2025. 2

[39] Dezhi Kong, Zhengzhao Feng, Qiliang Liang, Hao Wang, Haofei Sun, Changpeng Yang, Yang Li, Peng Zhou, Shuai Nie, Hongzhen Wang, Linfeng Zhou, Hao Jia, Jiaming Xu, Runyu Shi, and Ying Huang. ProactiveMobile: A comprehensive benchmark for boosting proactive intelligence on mobile devices. In *CVPR*, pages 27503–27513, 2026. 3

[40] Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Alireza Salemi, Ryan A. Rossi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, Nedim Lipka, Chien Van Nguyen, Thien Huu Nguyen, and Hamed Zamani. LongLaMP: A benchmark for personalized long-form text generation. *arXiv preprint arXiv:2407.11016*, 2024. 3

[41] Kuang-Huei Lee, Xinyun Chen, Hiroki Furuta, John Canny, and Ian Fischer. A Human-Inspired reading agent with gist memory of very long contexts. In *Proceedings of the International Conference on Machine Learning*, 2024. 3

[42] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *NeurIPS*, 2020. 3

[43] Jiahao Nick Li, Zhuohao Jerry Zhang, and Jiaju Ma. OmniQuery: Contextually augmenting captured multimodal memory to enable personal question answering. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2025. 2

[44] Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026. 3, 7

[45] Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. Search-o1: Agentic Search-Enhanced large reasoning models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2025. 2, 3

[46] Zhijia Liang, Jiaming Li, Weikai Chen, Yanhao Zhang, Haonan Lu, and Guanbin Li. OASIS: On-demand hierarchical event memory for streaming video reasoning. In *CVPR*, pages 2821–2831, 2026. 3

[47] Ruotong Liao, Meikel Kokowski, Xiaoxiao Sun, Yongqin Xian, Laura Bravo Sánchez, Shiye Su, Federico Tombari, Serena Yeung-Levy, and Volker Tresp. ProactEgoMem: Benchmarking proactive memory in streaming egocentric video. *arXiv preprint arXiv:2610.30089*, 2026. 2, 3

[48] Junming Lin, Zheng Fang, Chi Chen, Zihao Wan, Fuwen Luo, Peng Li, Yang Liu, and Maosong Sun. StreamingBench: Assessing the gap for MLLMs to achieve streaming video understanding. *arXiv preprint arXiv:2411.03628*, 2024. 3

[49] Jingyang Lin, Jialian Wu, Jiang Liu, Ximeng Sun, Ze Wang, Xiaodong Yu, Jiebo Luo, Zicheng Liu, and Emad Barsoum. VideoSeek: Long-horizon video agent with tool-guided seeking. In *CVPR*, pages 5465–5475, 2026. 3

[50] Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. Seeing, listening, remembering, and reasoning: A multimodal agent with Long-Term memory. In *ICLR*, 2026. 2

[51] Yaxi Lu, Shenzhi Yang, Cheng Qian, Guirong Chen, Qinyu Luo, Yesai Wu, Huadong Wang, Xin Cong, Zhong Zhang, Yankai Lin, et al. Proactive agent: Shifting LLM agents from reactive responses to active assistance. In *ICLR*, 2025. 2, 3

[52] Zhaoyang Lv, Nicholas Charron, Pierre Moulon, Alexander Gamino, Cheng Peng, Chris Sweeney, Edward Miller, Huixuan Tang, Jeff Meissner, Jing Dong, et al. Aria everyday activities dataset. *arXiv preprint arXiv:2402.13349*, 2024. 3

[53] Lingni Ma, Yuting Ye, Rowan Postyeni, Alexander J. Gamino, Vijay Baiyya, Luis Pesqueira, Kevin M. Bailey, David Soriano Fosas, Fangzhou Hong, Vladimir Guzov, et al. Nymeria: A massive collection of egocentric multimodal human motion in the wild. In *ECCV*, 2024. 3

[54] Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870, 2024. 2, 3

[55] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A diagnostic benchmark for very long-form video language understanding. In *NeurIPS*, 2023. 2, 3

[56] Abhinav Mehrotra, Veljko Pejovic, Jo Vermeulen, Robert Hendley, and Mirco Musolesi. My phone and me: Understanding people's receptivity to mobile notifications. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1021–1032, 2016. 3

[57] Jingbiao Mei, Jinghong Chen, Guangyu Yang, Xinyu Hou, Margaret Li, and Bill Byrne. According to me: Long-term personalized referential memory QA. *arXiv preprint arXiv:2603.01990*, 2026. 2, 3

[58] Thao Nguyen, Haotian Liu, Yuheng Li, Mu Cai, Utkarsh Ojha, and Yong Jae Lee. Yo'LLaVA: Your personalized language and vision assistant. In *NeurIPS*, 2024. 3

[59] Chang Nie, Chaoyou Fu, Yifan Zhang, Haihua Yang, and Caifeng Shan. PersonaVLM: Long-term personalized multimodal LLMs. In *CVPR*, pages 15000–15009, 2026. 2, 3

[60] Junbo Niu, Yifei Li, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, and Jiaqi Wang. OVO-Bench: How far is your Video-LLMs from Real-World online video understanding? In *CVPR*, pages 18902–18913, 2025. 3

[61] Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023. 2

[62] Renjie Pi, Jianshu Zhang, Tianyang Han, Jipeng Zhang, Rui Pan, and Tong Zhang. Personalized visual instruction tuning. In *ICLR*, 2025. 3

[63] Mitchell Piehl and Muchao Ye. MemFit: Efficient long-term agentic memory. *arXiv preprint arXiv:2610.00872*, 2026. 2, 3, 6, 7

[64] Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Dispider: Enabling video LLMs with active Real-Time interaction via disentangled perception, decision, and reaction. In *CVPR*, 2025. 3

[65] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, 2021. 3

[66] Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. Zep: A temporal knowledge graph architecture for agent memory. *arXiv preprint arXiv:2501.13956*, 2025. 2

[67] Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. LaMP: When large language models meet personalization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2024. 3

[68] Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. RAPTOR: Recursive abstractive processing for Tree-Organized retrieval. In *ICLR*, 2024. 3

[69] Soroush Seifi, Simon Gardier, Vaggelis Dorovatas, Daniel Olmeda Reino, and Rahaf Aljundi. Ego: Embedding-guided personalization of vision-language models. In *CVPR*, pages 11674–11683, 2026. 3

[70] Ashish Seth, Xinhao Mei, Changsheng Zhao, Varun Nagaraja, Ernie Chang, Gregory P. Meyer, Gael Le Lan,

Yunyang Xiong, Vikas Chandra, Yangyang Shi, Dinesh Manocha, and Zhipeng Cai. EgoAVU: Egocentric audio-visual understanding. In *CVPR*, pages 15805–15814, 2026. 3

[71] Yufei Shi, Weilong Yan, Gang Xu, Yumeng Li, Yucheng Chen, Zhenxi Li, Fei Richard Yu, Ming Li, and Si Yong Yeo. PVChat: Personalized video chat with One-Shot learning. In *ICCV*, 2025. 3

[72] Nikesh Subedi, Loris Bazzani, and Ziad Al-Halah. Interactive episodic memory with user feedback. In *CVPR*, pages 38826–38835, 2026. 3

[73] Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. Is ChatGPT good at search? investigating large language models as Re-Ranking agents. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2023. 3

[74] Ly-Duyen Tran, Manh-Duy Nguyen, Duc-Tien Dang-Nguyen, Silvan Heller, Florian Spiess, Jakub Lokoč, Ladislav Peška, Thao-Nhu Nguyen, Omar Shahbaz Khan, Aaron Duane, et al. Comparing interactive retrieval approaches at the lifelog search challenge 2021. *IEEE Access*, 2023. 3

[75] Wenhao Wang, Peizhi Niu, Gongyi Zou, Xiyuan Yang, Jingxing Wang, Haoting Shi, Yaxin Du, Jingyi Chai, Xianghe Pang, Shuo Tang, Yanfeng Wang, and Siheng Chen. MCP-Persona: Benchmarking LLM agents on real-world personal applications via environment simulation. In *International Conference on Machine Learning (ICML)*, 2026. 8

[76] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. HoloAssist: an egocentric human interaction dataset for interactive AI assistants in the real world. In *ICCV*, 2023. 3

[77] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. VideoAgent: Long-form video understanding with large language model as agent. In *ECCV*, 2024. 3

[78] Xiaodong Wang, Xuanyi Zhao, Pedro Rodriguez, Devendra Singh Sachan, Barlas Oguz, Seungwhan Moon, Shang-Wen Li, Gargi Ghosh, Xin Dong, and Wen-Tau Yih. S-EMBER: A large-scale benchmark for streaming egocentric memory retrieval. *arXiv preprint arXiv:2607.02689*, 2026. 3

[79] Yueqian Wang, Xiaojun Meng, Yuxuan Wang, Jianxin Liang, Jiansheng Wei, Huishuai Zhang, and Dongyan Zhao. VideoLLM knows when to speak: Enhancing Time-Sensitive video comprehension with Video-Text duet interaction format. In *Findings of the Association for Computational Linguistics: EMNLP*, 2025. 3

[80] Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. VideoTree: Adaptive tree-based video representation for LLM reasoning on long videos. In *CVPR*, 2025. 3

[81] Ziyang Wang, Yue Zhang, Shoubin Yu, Ce Zhang, Zengqi Zhao, Jaehong Yoon, Hyunji Lee, Gedas Bertasius, and Mohit Bansal. EgoMemReason: A memory-driven reasoning benchmark for long-horizon egocentric video understanding. In *Conference on Language Modeling (COLM)*, 2026. 3

[82] Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *ICLR*, 2025. 2, 7

[83] Junda Wu, Hanjia Lyu, Yu Xia, Zhehao Zhang, Joe Barrow, Ishita Kumar, Mehrnoosh Mirtaheri, Hongjie Chen, Ryan A. Rossi, Franck Dernoncourt, et al. Personalized multimodal large language models: A survey. *arXiv preprint arXiv:2412.02142*, 2024. 3

[84] Sean Wu, Aneesh Pappu, Jiaru Zou, Yubin Kim, Yuncong Mao, Hongjian Zhou, Junchi Yu, Jieneng Chen, and James Zou. Measuring AI agents as teammates with collaborative intelligence. *arXiv preprint arXiv:2611.02633*, 2026. 7

[85] Junbin Xiao, Shenglang Zhang, Pengxiang Zhu, and Angela Yao. Ego-grounding for personalized question-answering in egocentric videos. In *CVPR*, pages 40537–40547, 2026. 3

[86] Jing Xu, Arthur Szlam, and Jason Weston. Beyond goldfish memory: Long-Term Open-Domain conversation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2022. 2

[87] Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-MEM: Agentic memory for LLM agents. In *NeurIPS*, 2025. 2, 6, 7

[88] Yue Xu, Yutao Sun, Yihao Liu, Mengyu Zhou, Jiayi Qiao, Lu Ma, Kai Tang, Wenjie Wang, Xiaoxi Jiang, and Guanjun Jiang. From passive retrieval to active memory navigation: Learning to use memory as a structured action space. *arXiv preprint arXiv:2607.05794*, 2026. 3

[89] Jiaqi Yan, Ruilong Ren, Jingren Liu, Shuning Xu, Ling Wang, et al. TeleEgo: Benchmarking egocentric AI assistants in the wild. *arXiv preprint arXiv:2510.23981*, 2025. 3

[90] Bufang Yang, Lilin Xu, Liekang Zeng, Kaiwei Liu, Siyang Jiang, Wenrui Lu, Hongkai Chen, Xiaofan Jiang, Guoliang Xing, and Zhenyu Yan. ContextAgent: Context-aware proactive LLM agents with open-world sensory perceptions. In *NeurIPS*, 2025. 5

[91] Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, et al. EgoLife: Towards egocentric life assistant. In *CVPR*, 2025. 2, 3

[92] Jenny Yang, Elizabeth Proehl, Anjan Bharadwaj, Emily Fan, Charlotte Cole, James Kim, Noam Brown, Phoebe Thacker, Amelia Glaese, and Tejal Patwardhan. LifeVal: Evaluating AI model performance on personalized agentic consumer tasks. *arXiv preprint arXiv:2611.06687*, 2026. 3

[93] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *ICLR*, 2023. 2, 3

[94] Woongyeong Yeo, Kangsan Kim, Jaehong Yoon, and Sung Ju Hwang. WorldMM: Dynamic multimodal memory

agent for long video reasoning. In *CVPR*, pages 25599–25609, 2026.

[95] Xinlei Yin, Xiulian Peng, Xiao Li, Zhiwei Xiong, and Yan Lu. Hierarchical long video understanding with audiovisual entity cohesion and agentic search. In *CVPR*, pages 32882–32891, 2026.

[96] Yufei Yin, Qianke Meng, Minghao Chen, Jiajun Ding, Zhenwei Shao, and Zhou Yu. VideoARM: Agentic reasoning over hierarchical memory for long-form video understanding. In *CVPR*, pages 24042–24051, 2026. 3

[97] Shoubin Yu, Lei Shu, Antoine Yang, Yao Fu, Srinivas Sunkara, Maria Wang, Jindong Chen, Mohit Bansal, and Boqing Gong. Ego2Web: A web agent benchmark grounded in egocentric videos. In *CVPR*, pages 25633–25643, 2026. 3

[98] Zihao Yu, Xiu Yuan, and Chongjie Zhang. MEMORA: Embodied action memory from egocentric videos for reasoning and planning. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 25551–25600, 2026. 3

[99] Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple LLM framework for Long-Range video Question-Answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2024. 3

[100] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2018. 3

[101] Zeyu Zhang, Quanyu Dai, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model based agents. *ACM Transactions on Information Systems*, 43(6):1–47, 2025. 2

[102] Zhehao Zhang, Ryan A. Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, et al. Personalization of large language models: A survey. *Transactions on Machine Learning Research*, 2025. 3

[103] Qi Zhao, Shijie Wang, Ce Zhang, Changcheng Fu, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. AntGPT: Can large language models help long-term action anticipation from videos? In *ICLR*, 2024. 3

[104] Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. Do LLMs recognize your preferences? evaluating personalized preference following in LLMs. In *ICLR*, 2025. 3

[105] Jianjie Zheng, Zhichen Liu, Zhanyu Shen, Jingxiang Qu, Guanhua Chen, Yile Wang, Yang Xu, Yang Liu, and Sijie Cheng. Evaluating memory capability in continuous lifelog scenario. In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026. 2

[106] Yu Zhong, Tianwei Lin, Ruike Zhu, Yuqian Yuan, Haoyu Zheng, Liang Liang, Wenqiao Zhang, Feifei Shao, Haoyuan Li, Wanggui He, Hao Jiang, and Yueting Zhuang. Unified personalized understanding, generating and editing. In *CVPR*, pages 29400–29409, 2026. 3

[107] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. MLVU: Benchmarking multi-task long video understanding. In *CVPR*, 2025. 3